Automatikus információszerzés gazdasági rövidhírekből

Prószéky Gábor

MorphoLogic
Budapest
proszeky@morphologic.hu

Kivonat. Az alábbiakban bemutatjuk azt a magyar nyelvre készített információkivonatoló rendszert, a *NewsPro*-t, melyet az NKFP 2/017/2001 projekt keretében készített a MorphoLogic Kft., az MTA Nyelvtudományi Intézetének Korpusznyelvészeti Osztálya, a Szegedi Egyetem Informatikai Tanszékcsoportja és a Gallup Intézet alkotta konzorcium. A kutató-fejlesztő munkát a gazdasági, ezen belül is elsősorban a céginformációs rövidhírek információtartalmának automatikus kinyerésére összpontosítottuk.

1 Tartalomelemzés és információkinyerés

Az üzleti élet gyakorlatilag minden területén a legfontosabb információk elektronikus szövegek formájában állnak rendelkezésre. Az információs társadalom legfontosabb eszközévé az interneten elérhető szövegek váltak. A gazdasági, társadalmi, politikai élet legfontosabb döntéseihez szükséges információk hozzáférési modellje az elmúlt 5-6 évben gyökeresen megváltozott. Mivel a döntés-előkészítéshez szükséges legfontosabb információk egyre inkább csak a számítógépes szövegfeldolgozás segítségével érhetők el, a pályázatunkban megcélzott feladatok megoldása stratégiai fontosságúvá vált. A dolgok természetéből adódóan az új protokollok által hálózaton keresztül is elérhető szövegek és egyéb adatok döntő többsége angol nyelven állt elő. A legfontosabb szövegfeltáró technológiákat is – érthető módon, az angol nyelv hatalmas piaci potenciáljának megfelelően – erre a nyelvre fejlesztik. Ugyanakkor az internet használói között minden előrejelzés szerint a következő években mind a tartalom, mind a felhasználók nyelve tekintetében más nyelvek kerülnek többségbe. A projektünkben kifejlesztett eszközök létfontosságúak ahhoz, hogy a mindennapi elektronikus üzleti tranzakciók és más információalapú szolgáltatások a nemzeti nyelveken folyhassanak. Olyan kis nyelv esetében, mint a magyar, az ilyen feladat mielőbbi megoldásának multiplikátorhatása van: technikai megoldásokat nyújtunk arra, hogy az elektronikus szövegtengerből információt, illetve – tartalomelemzési eljárások kidolgozásával – adatbázisokba tölthető tudást konvertáljunk.

A fenti célok megvalósítása jelentős mértékben igényelt nyelvtechnológiai alapkutatást, mivel a magyar nyelvű szövegek számítógépes feldolgozásának lehetőségei – bár az elmúlt években jelentős fejlődést mutattak – a projekt kezdetekor nem álltak azon a fokon, hogy megfelelhessenek az információkinyerés igényeinek. Az információkinyerés (Information Extraction, IE) az 1990-es években, az internet elterjedésével és az elektronikusan rögzített szövegek robbanásának hatására lett az információtechnológiai alkalmazások egyik legfontosabb céljává és eszközévé vált [1] [2]. Fontosságának megértéséhez azonban meg kell különböztetnünk az információkigyűjtéstől (Information Retrieval, IR) és a számítógépes nyelvmegértéstől (Natural Language Understanding, NLU). Az információkigyűjtés célja olyan dokumentumok (vagy más adatok) kikeresése, amelyek illeszkednek egy adott lekér-

dezésre, viszont nem célja a tartalom strukturált ábrázolása. A nyelvmegértés adott szöveg teljes tartalmának számítógépes ábrázolására irányul. Az ilyen eszközök a legkülönbözőbb területeken működnek, ám csak olyan szövegeket tudnak teljesen feldolgozni, amelyek megfelelnek valamely egyszerűsített nyelvtannak, és nem tartalmaznak a nyelvtan számára ismeretlen nyelvi elemeket.

Az információkinyerés célja strukturált – gépileg lekérdezhető, feldolgozható – adathalmaz előállítása szöveges dokumentumok tartalmából. Az így létrejövő adatbázis nem a szövegeket, hanem a belőlük kinyert releváns adatokat tartalmazza. Az információkinyerésnek azonban nem célja a teljes szövegtartalom ábrázolása. Az információkinyerés során nagy mennyiségű szövegből gyűjtünk ki információt. A folyamat során először minden szövegben meg kell keresni a releváns információt, azt strukturált formában ki kell vonni, majd azt egy előre meghatározott struktúrában tárolni. Lényeges, hogy az eljárás figyelmen kívül hagyja a nem releváns információt. Ezek a lépések egyrészt szigorúbban meghatározzák a feladatot, mint természetesnyelv-feldolgozás többi területe. Az elkészült rendszerek hatékonyságát pedig könnyű tesztelni az emberi és a számítógépes teljesítmény összehasonlításával. Az – angol nyelvű rendszereken kapott – teszteredmények szerint a legtöbb esetben legalább 20-30%-kal több és pontosabb információt nyújt a számítógépes információkinyerő rendszer, és ehhez lényegesen kevesebb időt igényel, mint az ember. Az információkinyerő rendszerek gyakorlati felhasználásai között megtalálhatjuk az adatbázis-építést nem rendezett szövegekből (példák: szakirodalom követése, üzleti és gazdasági problémák nyomon követése a sajtó rövidhírei alapján), összefoglaló rendszerek és a trendek megállapítására használható adatbányászati alkalmazások alapadatainak, valamint az információkigyűjtő rendszerekben használt indexek létrehozását.

Az információkinyerő rendszerek nem törekszenek a szövegek teljes elemzésére és megértésére, csupán a releváns részeket emelik ki és elemzik a részleges mondatelemzés eljárásával. Ez egyrészt növeli a sebességet, másrészt – a gyakorlati tapasztalat alapján – ez elegendő is az információ kinyeréséhez. Az ilyen rendszerek többek közt ennek köszönhetően rugalmasabbak is, mint a nyelvmegértő rendszerek, hiszen összetett és ismeretlen – és esetleg rosszul formált, nyelvtanilag helytelen – mondatokon is működik meghatározott – korlátozott – szemantikai területeken.

A természetes nyelvek kutatása informatikai szempontból a modern számítógépek számára is komoly kihívást jelent mind háttértár-kapacitásban, mind feldolgozási sebességben. Ennek oka a természetes nyelvekben előforduló jelentős mennyiségű szóalak, illetve a beszélt és írott nyelvben használt változatos – formális eszközökkel nagyon nehezen követhető – mondatstruktúra. Tovább bonyolítja a helyzetet a természetes nyelvekben előforduló ismeretlen tulajdonnevek kezelése, amely további erőfeszítéseket követel, ugyanis a meglevő számítógépes nyelvleírások a szinkron modellt követik, vagyis a nyelv valamely pillanatnyi állapotát írják le, zárt szótáraikkal együtt. Az informatika és az internet elterjedése miatt a természetes nyelvek interferenciája is nagy. Jelentős számú idegen – másik természetes nyelvből származó – elem épül be az egyes nyelvi szövegekbe. A hatékony informatikai rendszerek egyszerre több nyelvre is alkalmazható általános reprezentációt kell használjanak a feldolgozás különböző szintjein.

A kutatási erőfeszítések többsége jelenleg is az angol nyelvre összpontosul. Az információreprezentálásra irányuló magyar nyelvi kutatásokban eddig legfeljebb csak a kezdeti lépések történtek meg. Konzorciumunk tagjai közül többen vettek részt európai projektekben is, és az ott kidolgozott általános módszereket alkalmazták a magyar nyelvre is. A konzorcium e projekt keretében a magyar nyelvvel kapcsolatosan egyaránt folytatott alap- és alkalmazott kutatásokat. Reményeink szerint a projekt hosszú távra meghatározza a téma további kutatási irányait, mert olyan – a kutatások keretében referenciaként használható – szintaktikai és szemantikai kódolási technológiák, korpuszábrázolási modellek, definíciók készültek,

amelyek közös platformot teremtettek a további hasonló munkákhoz. A konzorcium további tevékenysége a projektben a magyar szövegek tartalmi feldolgozását előkészítő alap- és alkalmazott kutatásokra irányult. A célul kitűzött információkinyerési technológia kifejlesztéséhez olyan mondatelemző rendszer volt szükséges, amely nem törekszik feltétlenül szöveg mondatainak teljes elemzésére, azonban alkalmas arra, hogy a szövegben elszórtan szereplő – formálisan definiálható – szerkezeti elemeket felismerje, rendezze, kivonatolja, azaz információt vonjon ki belőlük. Ehhez kapcsolódó alapkutatási feladat volt a magyar nyelvben előforduló egyes (pl. főnévi) szerkezetek struktúrájának, szintaxisának vizsgálata is. Az utóbbinak feltétele volt, hogy feltárják a magyar nyelv szókincsének morfo-szintaktikai és egyes – korlátozott jelentéskörben érvényes – szemantikai jegyeit is. Alapkutatásunk többek közt tehát erre is irányult.

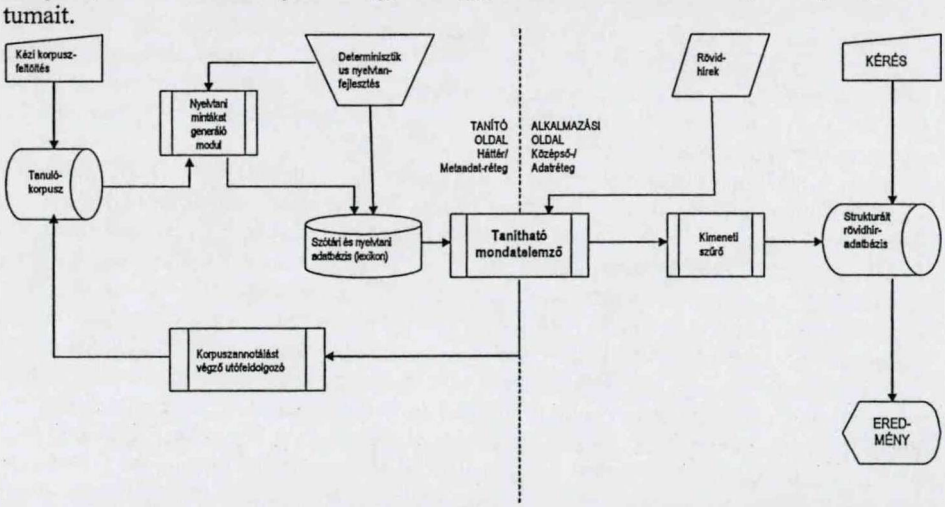
2 A kifejlesztett NewsPro rendszer

A *NewsPro* rendszer architektúrája az 1. ábrán látható. A kutató-fejlesztő munkát a gazdasági, üzleti, pénzügyi, piaci, ezen belül is elsősorban a céginformációs rövidhírek információtartalmának automatikus kinyerésére összpontosítottuk. Az összegyűjtött híreket egységes annotációval (tartalmi kódolással) láttuk el, a Reuters, az AFP vagy a UPI által is széles körűen használt NewsML-szabvány [3] szerint. Az első feladat olyan adatbázis létrehozása volt, amely a magyar nyelv törzsszókincsének minden eleméhez tartalmazza a sikeres nyelvi (elsősorban mondat- és szöveg-) elemzéshez szükséges morfológiai (szóalaktani), szintaktikai (szóhatáron túli nyelvtani) és szemantikai (jelentésleíró) információt. Az utóbbi esetben csak alapvető (szótári) információról lehet szó, hiszen a jelentés teljes ábrázolása az emberi tudás olyan részletességű ábrázolását jelentené, ami a tudomány mai állása szerint belátható időn belül nem valósítható meg. Az eredményül kapott adatbázis külön tartalmazza az igei vonzatkeretek, illetve a főnevek és a melléknevek leírását. Az adatbázis elemeit jól definiált jegyekkel láttuk el. Ezután kialakítottuk azt az infrastruktúrát, amely lehetővé teszi, hogy a program tetszőleges nyelvtant befogadjon, és ennek felhasználásával elvégezze szövegek részleges vagy teljes elemzését. A program semmilyen előfeltételezéssel nem él a nyelvtant illetően, vagyis teljesen nyelvfüggetlen, illetve adatvezérelt, egyedül a nyelvtan leírásának formátuma van megkötve.

A mondatelemzés alapjául szolgáló szabályrendszer (mintanyelvtan) kifejlesztéséhez megfelelően annotált gyakorlókorpusz szükséges, ezért elengedhetetlen fontosságú volt egy nyelvtani szempontból több szinten annotált korpusz kifejlesztése. A gyakorlókorpusz annotálása két szinten történt. Az annotáció első szintje az egyes szavak helyes szófaji kóddal való ellátását (morfológiai elemzését) foglalja magába. Ez az annotálás egy az SZTE és a MorphoLogic korábbi projektje (IKTA-027/2000) során kifejlesztett szófaji egyértelműsítő program segítségével történt [4]. A szófaji egyértelműsítés egyben a szintaktikai elemzés előfeltétele. Az elemzés második szintje a szavak, szószerkezetek mondatban betöltött szerepének meghatározását, ezen belül is a főnévi csoportok bejelölését, azaz szintaktikai annotálást jelenti. Az annotálást egy előelemző eszköz segítségével végeztük, majd ezt követően az erre a célra kifejlesztett eszköz felhasználásával félautomatikus módon, kézzel javítottuk. A fentiek alapján az annotált korpusz elemzési fák adatbázisának (treebank-nek) is tekinthető, s mint ilyen, a maga nemében az első a magyar nyelvterületen. Ez a munka előkészítése volt egy nagyobb, általános mondatelemzési adatbázisnak (IKTA-5/037/2002).

A mondatelemzésben a főnévi szerkezetek leírására kifejlesztendő szabályrendszer két pillérre épül. Egyrészt a mondatelemző program tanulási képességeinek felhasználásával, induktív logikai eljárásokkal tanuló-adatbázis alapján felismerési szabályokat állítottunk elő az egyes szintaktikai elemekhez. A program tanulási képességének rugalmassága abban

nyilvánul meg, hogy a mintanyelvtan működés közben folyamatosan bővíthető, mert a rendszer nem algoritmikus szabályokra, hanem mintaillesztésre épül. A másik módszer a hagyományos nyelvészeti kutatás volt, melynek keretében meghatároztuk a magyar nyelv olyan mondatelemeit, amelyek relevánsak a kifejlesztendő tartalomelemzési technológia szempontjából. Már a csak morfoszintaktikai (kb. szófaji) kódokkal ellátott korpuszban is alkalmazhatók egyszerű mintaillesztési algoritmusok, reguláris kifejezések segítségével írt nyelvtanok, melyek alapján a különböző mondatszerkezeti elemek azonosíthatók. Az annotált korpusz elkészítésének az volt a célja, hogy azt feldolgozva különféle gépi tanulási technikák alkalmazásával olyan tudásbázist állítsunk elő, ami megfelelő minőségű támogatást biztosít egy automatikus szintaktikai elemző számára. Az eredmények szintaktikai szabályok, amelyek a HumorESK mondatelemző program [5] meglevő szabályai közé megfelelő konverzióval beilleszthetők. Műhelyünkben végül is *HumorESK LangXtract* néven olyan információkivonatoló program keletkezett, amely a korábban kifejlesztett HumorESK mondatelemző programot felhasználva a bemeneti szövegben megjelöl meghatározott nyelvi elemeket. A mondatelemző program rendkívül részletes információt ad egy-egy szövegrész nyelvtani szerkezetéről. A nyelvi kivonatoló program szabadon konfigurálható a tekintetben, hogy az elemzőmodul által visszaadott információtömegből mely elemeket értékeljük relevánsnak. Így e program segítségével a tartalomelemzés szempontjából érdekes nyelvi elemeket úgy lehet kiemelni, hogy a feldolgozást nem zavarja az elemzés során létrehozott többi, a nyelvi struktúra köztes szintjeit felépítő szimbólum – vagyis a program elvégzi a válogatás jelentős részét. Ennek folytatásaként elkészült a *FrameTagger* nevű program prototípusa, mely egyrészt elvégzi a szövegben a HumorESK LangXtract által megjelölt névszói és más nyelvi szerkezetek szemantikai annotációját, továbbá absztrakt eseménymintákat (szemantikai kereteket) próbál meg a mondatokra illeszteni. Sikeres illeszkedés esetén automatikusan meghatározza az adott esemény szereplőit, körülményeit, attribútumait. A program az input állományban azokat a mondatokat, amelyekre sikeres illesztést tudott végrehajtani, XML-címkékkel jelöli meg, csakúgy, mint a felismert esemény szereplőit, attribútumait.



1. ábra. A rövidhír-feldolgozó NewsPro rendszer architektúrájának vázlata

3 Egy példa a NewsPro rendszer alkalmazására

Az elkészített prototípus-rendszer egy nagyobb témakör, az üzleti, azon belül is a céginformációs rövidhírek világában igazodik el elsősorban. Az elkészített prototípus fő célja annak megmutatása volt, hogy a kitűzött célt, az üzleti hírekben szereplő egyszerűbb események automatikus felismerését és a hírekben szereplő információk automatikus kinyerését meg lehet valósítani. Így tehát attól a szövegtől, hogy

Az Erste Bank 16,5 százalékkal növelte nyereségét 164,6 millió euróról 191,8 millió euróra.

Ennek NewsPro-feldolgozása pedig valahogy így fest XML-ben:

```
<root id="640" class="S-FULL" start="1" end="14">
<event schema="increased.midterm_report.income.1" roles_matched="6/6">
- <rv role="company" pos="N" case="NOM" sem="company|institute">
- <NP id="85" class="NP-FULL" start="1" end="3" head_lex="bank"
      HSK_head_lex="bank" case="NOM" ownernum="nil" ownerpers="nil"
      postp="" sem="company countable human institute">
    <w id="0" class="DET" at="1-1" lex="az" case="NOM">Az</w>
    <w id="2" class="UNKNOWN" at="2-2" lex="Erste">Erste</w>
    <w id="4" class="N" at="3-3" lex="bank" case="NOM">Bank</w>
  </NP>
  </rv>
- <rv role="_1" pos="V" lemma="növel">
    <w id="10" class="V" at="6-6" lex="növel">növelte</w>
  </rv>
- <rv role="trade" pos="N" lemma="eredmény|nyereség|profit" case="ACC"
      possessed_by="company" sem="abstract">
- <NP id="107" class="NP-FULL" start="7" end="7" head_lex="nyereség"
      HSK_head_lex="nyereség" case="ACC" ownernum="nil" ownerpers="nil"
      postp="" sem="abstract">
    <w id="12" class="N" at="7-7" lex="nyereség" case="ACC">nyereségét</w>
  </NP>
  </rv>
- <rv role="measure" pos="N" lemma="százalék" case="INS"
      modified_by_number="YES">
- <NP id="97" class="NP-FULL" start="4" end="5" head_lex="százalék"
      HSK_head_lex="16,5" case="INS" ownernum="nil" ownerpers="nil"
      postp="" sem="abstract countable">
    <w id="6" class="NUM" at="4-4" lex="16,5" case="NOM">16,5</w>
    <w id="8" class="N" at="5-5" lex="százalék" case="INS">százalékkal</w>
  </NP>
  </rv>
- <rv role="old_value" pos="N" case="DEL" modified_by_number="YES"
      sem="currency">
- <NP id="122" class="NP-FULL" start="8" end="10" head_lex="euró"
      HSK_head_lex="164,6" case="DEL" ownernum="nil" ownerpers="nil"
      postp="" sem="countable currency measure">
    <w id="15" class="NUM" at="8-8" lex="164,6" case="NOM">164,6</w>
    <w id="17" class="NUM" at="9-9" lex="millió" case="NOM">millió</w>
    <w id="19" class="N" at="10-10" lex="euró" case="DEL">euróról</w>
  </NP>
  </rv>
- <rv role="new_value" pos="N" case="SUB" modified_by_number="YES"
      sem="currency">
- <NP id="137" class="NP-FULL" start="11" end="13" head_lex="euró"
      HSK_head_lex="191,8" case="SUB" ownernum="nil" ownerpers="nil"
      postp="" sem="countable currency measure">
    <w id="21" class="NUM" at="11-11" lex="191,8" case="NOM">191,8</w>
    <w id="23" class="NUM" at="12-12" lex="millió" case="NOM">millió</w>
    <w id="25" class="N" at="13-13" lex="euró" case="SUB">euróra.</w>
  </NP>
  </rv>
</event>
```

Ez a kimenet nem olvasható könnyen „emberi" szemmel, de annyit észrevehetünk, hogy az információkivonatolás a nyelvtani elemzés eredményeire épül. A NewsPro rendszer a mondatban úgynevezett eseménysémákat (event schema) azonosított. Az eseménysséma meghatározza az esemény fajtáját és a résztvevőket. Ha tehát egy vállalat tulajdonosváltásáról van szó, akkor a rendszer „tudja", hogy itt meg kell nevezni a vevőt, az eladót, az adásvétel tárgyát, az árat (ha rendelkezésre áll), illetve a kérdéses tulajdoni hányadot. Ezek az elemzésben az „rv role" kezdetű sorokban láthatók. E struktúrát – vagy szűrés után egyes elemeit – kell adatbázisban tárolni. Ezt megkönnyíti, hogy a rendszer szabványos témabesorolást és esemény-, illetve szerep-megnevezéseket alkalmaz, a kimenet pedig a NewsML-szabványt követi. A könnyebb áttekinthetőség kedvéért azonban létrehoztuk a rendszer relációsadatbázis-szerű kimenetét is, melyet a fenti példán illusztrálva bemutatunk:

1. **increased.midterm_report.income.1** (6/6)

company	Erste Bank
_1	növel
trade	nyereség
measure	16,5 százalék
old_value	164,6 millió euró
new_value	191,8 millió euró

4 További fejlesztések, alkalmazások

Miután a NewsPro rendszert erre a célra kifejlesztett referenciakorpuszon validáltuk, ellenőriztük, a projekt végső eredménye egy olyan, kompakt, más rendszerekbe ágyazható keretrendszer lett, amely – megfelelő adatbázissal feltöltve – alkalmas arra, hogy rövid szöveges adatok tartalomelemzését felhasználó további alkalmazásokban is helyt álljon. Ahogyan az internet tovább terjed, a szöveges formában előálló adattömeg is várhatóan exponenciálisan növekszik tovább. A projekt által előállított megoldásnak a lényege részben az, hogy az intelligens tartalomelemzés – meghatározott területeknek megfelelő – szótárainak és résznyelvtanainak segítségével az egyes szövegek leírhatók a kötött, attribútumokkal rendelkező adatbázis-modell alapján. Az így „kitöltött" adatbázis-mezők azután a legkülönbözőbb kvalitatív, tér- és időbeli attribútumok mentén is lekérdezhetőkké válhatnak.

5 Hivatkozások

1. Cowie, J and Lehnert, W. Information Extraction. *Communications of the ACM*, 39(1) (1996)
2. Gaizauskas, R and Wilks, Y. Information Extraction: Beyond Document Retrieval. *Journal of Documentation* 54/1 (1998)
3. *Introduction to NewsML*: `www.newsml.org` (2003)
4. Alexin, Z. – J. Csirik – T. Gyimóthy – K. Bibok – Cs.Hatvani – G. Prószéky – L. Tihanyi. Manually Annotated Hungarian Corpus. In: Paroubek, P. (ed.) *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Vol.2., 53—56, Budapest (2003)
5. Prószéky, G.: Syntax As Meta-morphology. *Proceedings of COLING-96,* Vol.2, 1123–1126. Copenhagen, Denmark (1996)